

Your Second RecSys

Дополнительные методы оценки качества рекомендаций

Михаил Хасыков

ML-разработчик

MTC BigData



План

Визуальный анализ

- Способы получить качественную оценку
- Как проводить визуальный анализ

Beyond accuracy metrics

- Зачем рекомендациям свойства разнообразия и новизны
- Метрики diversity/novelty/serendipity
- Как растить/балансировать метрики

Визуальный анализ

Качественная оценка

При аб-тестировании и оценке оффлайн-метрик
не получаем качественную оценку качества
рекомендаций

Зачем?

- Польза для **интерпретации количественных оценок**
- Помощь в **генерации гипотез** что и как улучшить

Что делать?

- Спросить пользователей
- User studies/experiments
- Ну или просто “**посмотреть глазами**” на рекомендации

Визуальный анализ

Метод получения **субъективной качественной оценки** рекомендательной системы путем “осмотра глазами” результатов ее работы разработчиком

Смотрим на рекомендации:

- Аватаров
- Реальных пользователей

Аватар/Персона

вымышленный человек, объединяющий в себе набор характеристик, присущих сегменту целевой аудитории, или проще – собирательный образ пользователя

применяются в ориентированных на пользователя сервисах и маркетинге

[Falk K., Practical Recommender Systems \(2019\)](#)

Therese Comedy		Helle Action	
Anything that makes me laugh is a hit.	User ID: 400003	Anything with superheroes and anything that blows up	User ID: 400004
Pietro Drama		Ekaterina Drama, action, comedy	
The more complicated the drama, the better it is.	User ID: 400005	Nothing beats drama, but I sometimes watch action, and rarely a comedy.	User ID: 400006

Как проводить визуальный анализ

На что обращаем внимание:

- **признаки** из профиля пользователя
- **логи действий в обучающей выборке**, их агрегаты
- **логи действий в тестовой выборке**, в том числе контекст
- **товары в рекомендациях**

На какие вопросы пытаемся ответить:

- **релевантны** ли рекомендации пользователю?
- как можно **сделать** рекомендации еще **лучше**?

Пример

Профиль пользователя

sex	age	kids_flg
M	age_45_54	0

Просмотры в трейне

title	genres	release_year
Терминатор	боевики, фантастика, триллеры	1984.0
Форрест Гамп	драмы, мелодрамы	1994.0

Просмотры в тесте

title	genres	release_year
Король лев (2019)	драмы, мультфильм, мюзикльные	2019.0

Рекомендации

title	genres	release_year
Робокоп	боевики, фантастика, триллеры	1987.0
Человек дождя	драмы	1988.0
Хищник	боевики, фантастика, триллеры, приключения	1987.0
Терминал	драмы, комедии	2004.0

Наблюдения:

- пользователь смотрел классические фильмы из 80-90х
- порекомендовали классические фильмы из тех же жанров, но не угадали

Гипотезы:

- под профилем совершен просмотр другим человеком?
- интересы пользователя разнообразны

Выводы:

- в целом, рекомендации “не плохи”
- может быть будет полезным большее разнообразие по жанрам и году релиза

Как использовать результаты

Аватары - проверка, что справляемся с “игрушечным” примером

Реальные пользователи:

- генерация гипотез о том, **что улучшить** в рекомендациях
- **интерпретация оценок оффлайн и онлайн-метрик:** выдвигаем гипотезы почему один алгоритм выиграл, а другой - проиграл
- в некоторых случаях **аргумент в решении о старте аб-теста**

Beyond accuracy metrics

Чего не хватает этим рекомендациям?



Разнообразие и новизна

- снижаем риск не угадать предпочтения пользователя
- боремся с формированием информационного пузыря
- увеличиваем продажи в длинном хвосте

Не во всех полках может быть уместно:

- похожие товары
- недавно просмотренные
- “часто покупаемые вами”

Похожие товары



Стиральная машина
ELECTROLUX EW6S3R26S



Стиральная машина CANDY
CSWS4 3642DB/2-07, с сушкой

Обозначения

i, j — индексы товаров

u, v — индексы пользователей

\mathcal{I}, \mathcal{U} — множества всех товаров и пользователей соответственно

\mathcal{I}_u — множество товаров, с которыми взаимодействовал пользователь u

\mathcal{U}_i — множество пользователей, которые взаимодействовали с товаром i

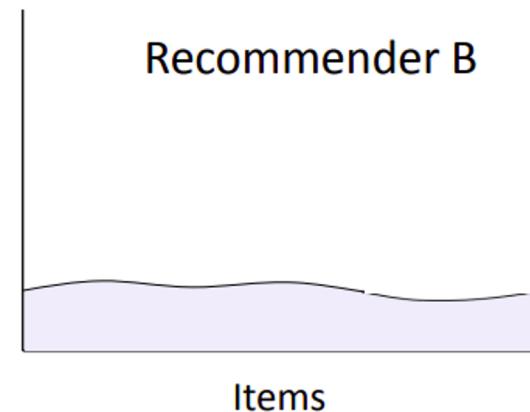
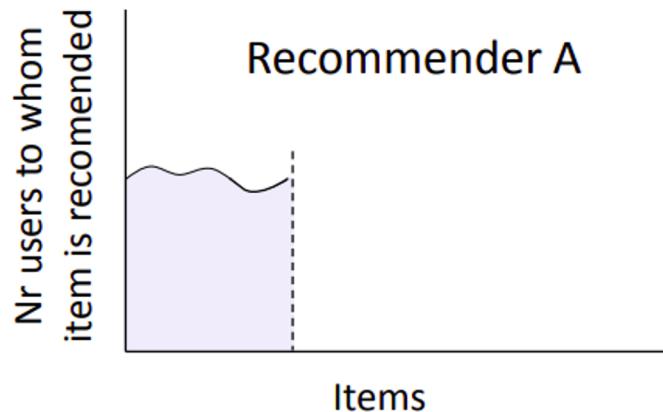
R_u — множество товаров, рекомендованных пользователю u

R — сокращение R_u в тех случаях, когда нет необходимости указывать индекс конкретного пользователя

Общее разнообразие

Aggregate diversity - количество товаров, в совокупности рекомендуемых алгоритмом всем пользователям (значения от 0 до $|\mathcal{I}|$)

$$\text{Aggdiv} = \left| \bigcup_{u \in \mathcal{U}} R_u \right|$$



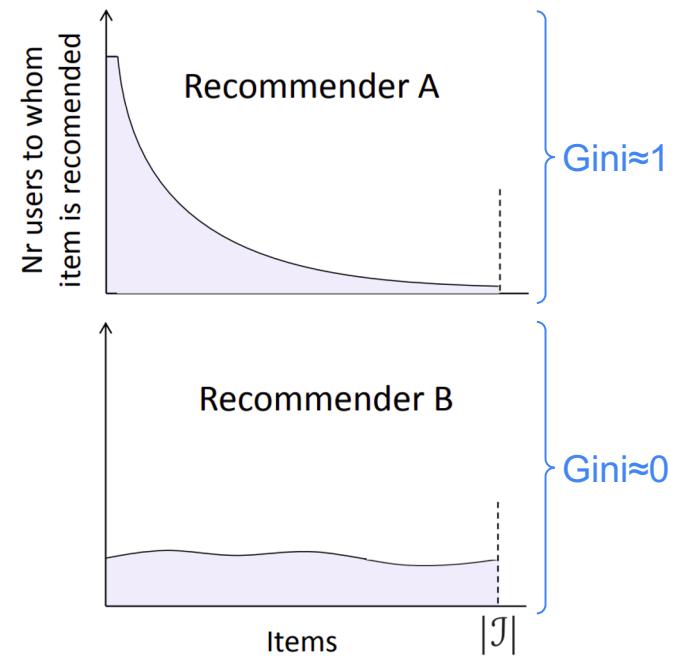
Коэффициент Gini

Gini измеряет равномерность распределения присутствия товаров в рекомендациях (значения от 0 до 1)

$$\text{Gini} = \frac{1}{|\mathcal{I}| - 1} \sum_{k=1}^{|\mathcal{I}|} (2k - |\mathcal{I}| - 1)p(i_k)$$

i_1, \dots, i_n - список товаров, отсортированный по вероятности вытягивания товара из списка рекомендаций:

$$p(i) = \frac{|\{u \in \mathcal{U} \mid i \in R_u\}|}{\sum_{j \in \mathcal{I}} |\{u \in \mathcal{U} \mid j \in R_u\}|}$$



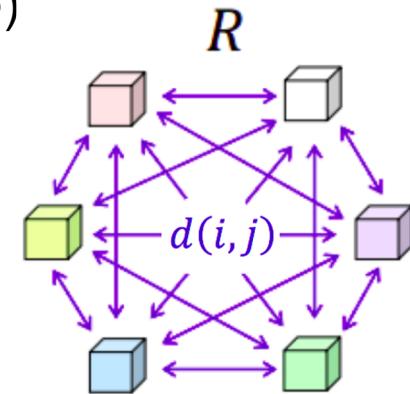
Разнообразие внутри полки

Intra-List Diversity определяется как среднее попарное расстояние между товарами в полке (значения от 0 до $+\infty$)

$$\text{ILD} = \frac{1}{|R|(|R| - 1)} \sum_{i \in R} \sum_{j \in R} d(i, j)$$

$d(i, j)$ определяются как расстояния между эмбеддингами или one-hot векторами признаков товаров:

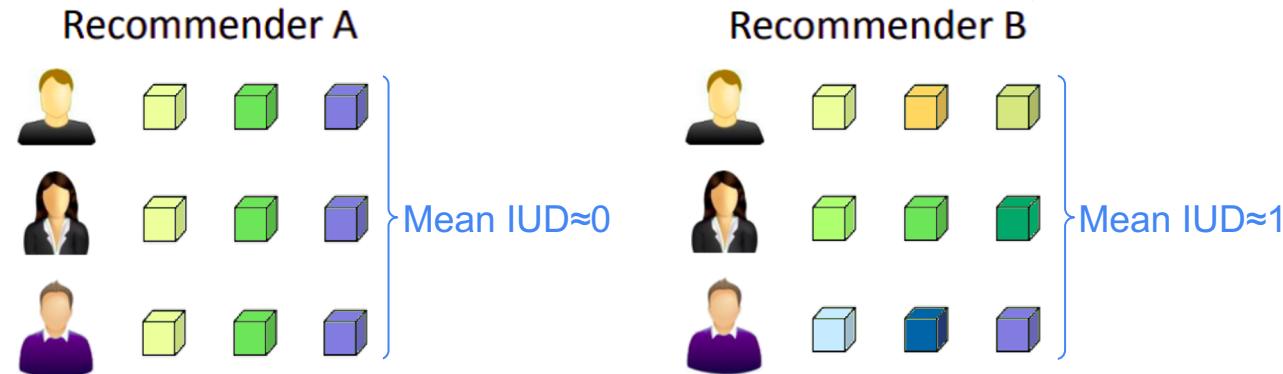
- евклидово
- косинусное
- Хэмминга и др.



“Персональность” рекомендаций

Inter-User Diversity - средняя доля пересечения рекомендаций для пользователя с рекомендациями для остальных пользователей (значения от 0 до 1)

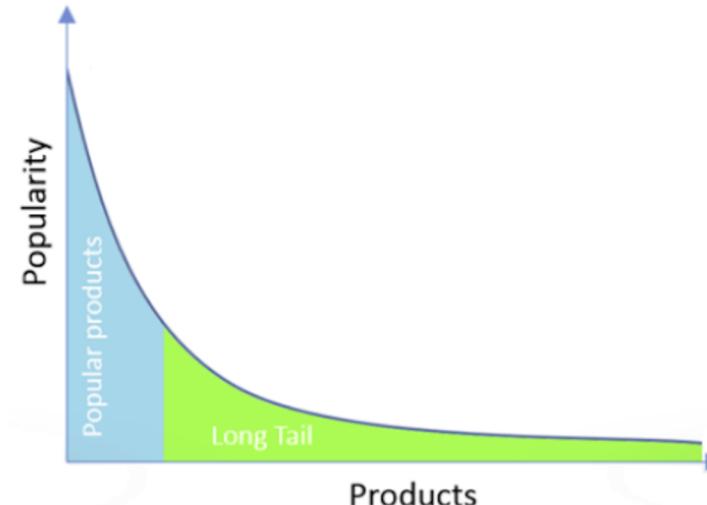
$$\text{IUD} = \frac{1}{|\mathcal{U}| - 1} \sum_{v \in \mathcal{U}} |R - R_v| / |R|$$



Новизна глобальная

Mean Inverse User Frequency - средняя новизна товаров в полке, где “новизна” товара обратно пропорциональна количеству пользователей, которые с ним взаимодействовали (значения от 0 до $+\infty$)

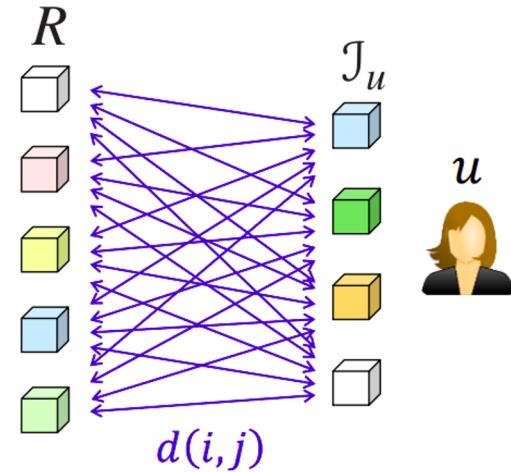
$$\text{MIUF} = -\frac{1}{|R|} \sum_{i \in R} \log_2 \underbrace{\frac{|\mathcal{U}_i|}{|\mathcal{U}|}}_{\text{популярность товара}}$$



Новизна персональная

Unexpectedness - отличие рекомендованных товаров от тех, с которыми пользователь уже взаимодействовал ранее (значения от 0 до $+\infty$)

$$\text{Unexp} = \frac{1}{|R||\mathcal{I}_u|} \sum_{i \in R} \sum_{j \in \mathcal{I}_u} d(i, j)$$



Serendipity (новизна + релевантность)



новизна



релевантность



Serendipity*

$$Serendipity = \frac{1}{|R|} \sum_{i \in R} \max(P_i - P_i^U, 0) \cdot rel_i$$

$$P_i = \frac{|\mathcal{I}| + 1 - rank_i}{|\mathcal{I}|}; P_i^U = \frac{|\mathcal{I}| + 1 - rank_i^U}{|\mathcal{I}|}$$

P_i, P_i^U — "вероятности" порекомендовать товар i пользователю и "произвольному" пользователю соответственно

$rank_i$ — ранг товара i в рекомендациях пользователю

$rank_i^U$ — ранг товара i в рекомендациях "произвольному" пользователю, например, можно определять как ранг по популярности

rel_i — мера релевантности товара i для пользователя

* по мотивам [Anna B. Recommender Systems — It's Not All About the Accuracy \(2016\)](#)

Методы увеличения метрик

изменение поведения базового алгоритма ранжирования

- например, изменение стратегии сэмплирования при оптимизации warp или bpr-лосса ([Jannach 2015](#))

постпроцессинг результатов базового алгоритма ранжирования (применимо к произвольному алгоритму)

- формирование полки с помощью оптимизации взвешенной комбинации меры релевантности и beyond accuracy метрики ([Ziegler 2005](#))

$$g(R, \lambda) = (1 - \lambda) \frac{1}{|R|} \sum_{i \in R} f_{rel}(i) + \lambda \operatorname{div}(R)$$

- наполнение полки товарами из разных кластеров товаров ([Li 2012](#))
- еще много методов по ссылкам на слайде “Ссылки”

Баланс между метриками

Рост метрик новизны и разнообразия обычно приводит к падению метрик **точности***

Как найти баланс?

- см. предыдущую лекцию про связь оффлайн и онлайн метрик
- формулируем продуктовую гипотезу о приоритете одной из метрик и увеличиваем ее несмотря на падение остальных метрик

*Good N. et al, Combining Collaborative Filtering with Personal Agents for Better Recommendations (1999)

Итоги

Визуальный анализ

- способ получения **качественной оценки** качества рекомендаций
- проводится путем **анализа рекомендаций** для аватаров и реальных пользователей

Разнообразие и новизна

- полезно оценивать и оптимизировать
- рассмотрели методы оценки с примерами
- поговорили о том, как увеличивать метрики и искать между ними баланс

Ссылки

Что почитать:

- [Ricci F. et al., Recommender Systems Handbook \(2011\) Chapter 26](#)
- [Berkovsky S. et al., Collaborative Recommendations: Algorithms, Practical Challenges and Applications \(2019\) Chapter 10](#)
- [Castells P., Recommender Systems Evaluation Beyond Accuracy \(2019\)](#)
- [Kaminskas M. et al., Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems](#)

Контакты

Михаил Хасыков

 mkhasykov@gmail.com

 [@mkhasykov](https://t.me/mkhasykov)

 <https://www.linkedin.com/in/mkhasykov>